Abstract

STRUCTURES TO REPRESENT POORLY FORMED HTML DOCUMENTS

Disclosed is a method of restructuring an input HTML document to comply with strict HTML. An input HTML document is linearly traversed to create a hierarchical tree structure representation (Figs. 2A-2F), the traversal maintaining a current insertion point (206, 210) for elements within the tree structure representation. During the traversal, elements (208) of the input HTML document that violate strict HTML are identified Each element is then processed individually, initially by retracing the tree structure representation from the current insertion point to identify an further insertion point from which the identified element can depend, the retracing comprising noting each parent element of the identified element passed during said retracing. Then, at the further insertion point, new elements (218) are created in the tree structure representation to correspond to those parent elements passed during the retracing, the new elements being created in reverse chronological order to that encountered during the retracing. The identified element (208) is then append to a terminal one of the new elements. The tree structure representation can then be converted into an output HTML document.